Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 24 (2013) 121 – 125

Procedia Computer Science

17th Asia Pacific Symposium on Intelligent and Evolutionary Systems, IES2013

Bias Reduction of Probabilistic Prototype Tree Based Estimation of Distribution Genetic Programming in Predicting Arthritis Prevalence

Kangil Kim[a,*], Hanggjun Cho[b]

[a] *Seoul National University*
Structural Complexity Laboratory
Building 302, Gwanangno 599
Seoul 151–744, Korea
[b] *Seoul National University*
Structural Complexity Laboratory
Building 302, Gwanangno 599
Seoul 151–744, Korea

Abstract

Estimation of Distribution Algorithms in Genetic Programming (EDA-GP) are algorithms applying stochastic model learning to genetic programming. In spite of various potential benefits, probabilistic prototype tree (PPT) based EDA-GPs recently appeared to have a critical problem of losing diversity easily. As an alternative learning method to reduce the effect, likelihood weighting (LW) was proposed and its results were positive to improve EDA-GP performance. In this paper, we aim to provide more generalised verification results to confirm the effects of LW. We investigate performance of PPT-based EDA-GP in a large scale problem predicting arthritis using medical data.

Selection and peer-review under responsibility of the Program Committee of IES2013

Keywords: Estimation of Distribution Algorithms, Genetic Programming, Probabilistic Prototype Tree, Arthritis, Disease Prediction, Bias

1. Introduction

In evolutionary computation, there has been a surge of research to apply stochastic models into genetic algorithms (GA), called Estimation of Distribution Algorithms (EDA)[?]. It was successful in improving performance and provided useful tools to incorporate knowledge. This approach is also applied to genetic programming (GP), called Estimation of Distribution Algorithms in Genetic Programming (EDA-GP)[?].

In EDA-GPs, however, complex data representation such as trees complicates building a stochastic model. This complexity led to developments of a variety of model representations so far. Recently, it was reported that a main representation, Probabilistic Prototype Tree (PPT)[?], may have critical problems by imposed bias[?]. According to this work, a model on PPT learns probability distribution significantly different from the model at previous generation in

* Corresponding author. Tel.: +82-2-860-6179
E-mail address: kangil.kim.01@gmail.com

Selection and peer-review under responsibility of the Program Committee of IES2013
doi:10.1016/j.procs.2013.10.034

process of using learning and sampling only. This bias is repeated at every generation and increases exponentially in terms of size of the model, which results in convergence to wrong solutions. To reduce this negative effect, modified mechanisms were proposed such as likelihood weighting. The method reduced the effect successfully and improved the performance of the EDA-GP systems.

In this paper, we aim to provide more practical evidence to support this argument, because the preliminary work is tested in well-known, but small-scale benchmark problems. For this verification, we will introduce a practical problem to predict a disease, arthritis, using medical data and then investigate the difference after reducing bias of a PPT-based EDA-GP. Arthritis is a well-known and commonly occuring disease, so predicting it is expected to have high impact for warning people and urging to manage their lifestyles and therapies. This problem is defined over partial data of National Health And Nutrition Examination Survey (NHANES) conducted by Centers of Disease Control and Prevention, USA[?].

The rest of this paper is organised as follows. Section 2 explains related background knowledge of EDA-GP, its bias problem and arthritis prediction. Section 3 describes how we define the prediction problem. Section 4 shows how to transform it to an optimisation problem to apply EDA-GP systems, and their configuration for experiments. Section 5 illustrates experiment results and analysis and we will make conclusions at Section 6.

2. Related Works

2.1. Estimation of Distribution Algorithms in Genetic Programming

Estimation of Distribution Algorithms in Genetic Programming is an algorithm of applying a stochastic model into genetic programming[?]. We may see this algorithm as an iterative learning algorithm or an evolutionary search algorithm guided by a stochastic model. The basic process of this algorithm is equal to EDA, which is repetition of generating samples from a model, selecting more fit samples, and updating the model. Detailed process is shown in algorithm 1. Replacing crossover or mutation with model learning and sampling process, this algorithm can increase

Algorithm 1 Basic Process of EDA and EDA-GP

```
Initialising Population – generating individuals from an initial probability model
while not (condition for termination) do
    Fitness Evaluation – evaluate fitness of individuals through given fitness function
    Selection – select the best individuals in terms of the fitness
    Update – modify probability or structure of the stochastic model using the best individuals
    Sampling – generate new individuals from probability stored in the model
end while
```

performance of conventional evolutionary algorithms. Moreover, its explicit representation has been expected to be benefitial to analysis of intermediate behaviour of the algorithms and guiding search incorporating prior knowledge.

Compared to EDAs adapting well-founded model representation such as Bayesian networks, EDA-GP is complication in learning a model because its data such as trees includes more information to control than linear chromosomes. Main distinctive features of the data are the number of variables and structural constraints imposed to symbols[?]. In EDA-GP, to control the aspects, a variety of model representations have been proposed such as PPT or Stochastic Grammars[?]. PPT is a major representation introduced in Probabilistic Incremental Program Evolution (PIPE) used in various EDA-GP models[????]. It may be regarded as a simple Bayesian network without any dependency between variables, but PPT has more restriction on selecting values by the constraints. An example of PPT is depicted in figure 1. PPT is basically a tree, but each node is a random variable representating a multinomial distribution over symbols. When we observe a tree, it is overlapped onto the PPT and its symbols are counted as samples for matching variables. From many observed trees, the PPT model learns new probability distribution for each variable. Generating samples is usually ancestral sampling started from the root variable of PPT. In this paper, we will mainly discuss about this PPT-based models.

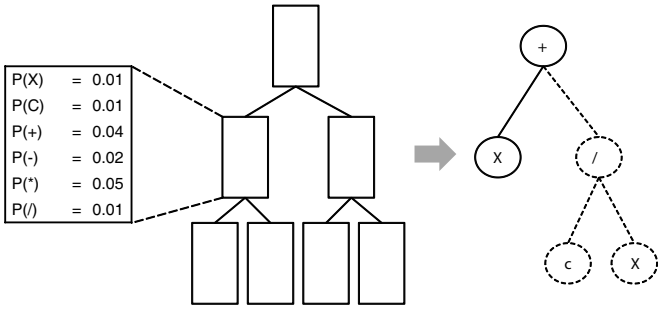


Fig. 1. A Probabilistic Prototype Tree Model Representation

2.2. Bias Reduction

In PPT-based models, bias can be generated in learning and sampling process. If we omit the selection from the EDA process, the loop is a simple process of generating samples from the model and learn the model from them again. Ordinary models composed of random variables is expected to learn the same distribution after the process, but PPT-based models show large difference. This bias is shown as a form of drift. Drift is not a significant factor changing distribution, but PPT representation increase the effect of nodes exponentially with respent to their depth. This phenomenon results in the convergence of the distribution to a point impossible to build optima[?].

In the same work, a likelihood weighting method is proposed to modify models for counteracting to the drft increase. Its basic idea is to use probability distribution of previous models for building new models, because probabilities of un-sampled proportion of total sample size is expected to be equal the probabilities of the models before sampling. This method is tested in various benchmark problems such as Max, partiy, and their variants and shown to improve performance and cancel bias in selection-omitted environment.

2.3. Arthritis

Arthritis is a disease involving inflammation of joints. It is the most common disease in the USA, showing 43.6% prevalence among 50 million adults (at 2002)[?]. This disease is categorised into various forms of arthritis such as Rheumatoid or Gout, and research to predict them has been heavily studied in medicine field[???], which implies the impact of arthritis prediction.

3. Prediction Problem of Arthritis Prevalence

In this paper, we define a problem to predict the prevalence of arthritis from simple body conditions of patients. It is defined over pre-processed partial data of NHANES callected from questionnaires of patients over all ages and races between 2007 and 2008.

From this data, we selected input parameters related to body conditions possibly affecting arthritis and the outcome representing prevalence of the disease, which is shown in the table 1. Used patient sample size is 4,738 without any missing value.

4. PPT-based EDA-GP for Arthritis Prediction

In the application of PPT-based EDA-GP for the arthritis problem, our primary aim is to investigate the difference of the system when LW method is applied. Secondary objective is to find an arithmetic function for prediction of prevalence, evaluating its accuracy through the given data. This empirical test is performed by designing solution space and EDA-GP system setting as follows.

Table 1. Features used for Arthritis Prediction

Parameter Type	ID(NHANES)	Description	Data Range
Input X_1	WHD010	current self-reported height	53–81
X_2	WHD020	current self-reported weight	83–440
X_3	WHD050	self-reported weight a year ago	78–440
X_4	PEASCTM1	blood pressure time in second	45–1521
X_5	PEASCTM2(IMQ020)	received hepatitis B 3 does series	23.2–55.5
X_6	BMXLEG	upper leg length	23.2–55.5
X_7	BMXWAIST	waist circumference	61–178.2
Output	MCQ160A	doctor ever said you had arthritis	0 or 1

Genotype Representation. The arithmetic functions are represented as tree individuals determined by a given symbol set. We set terminal symbols to $X_1, \ldots, X_7$ indicating input variables and used the ephemeral random constant (ERC)[?]. Operators are fixed as $+, -, \times, /$.

Fitness Evaluation. A fitness function to find the most accurate individuals is root mean square error (RMSE), which is a metric to evaluate difference between data for supervision and predicted outcome. In calculation of this error, the range of outcome is a set of all real number whereas the data is boolean value represented by 0 or 1. To map outcomes in the large range to boolean value, we adapted sigmoid function to scale the range of outcome. In applying PPT-based EDA-GP systems, shorter tree individuals are observed more than longer trees. This parsimony pressure is intentionally used in GP systems[??], but we set fitness function not to be biased by adding tree depth since the implicit bias of EDA-GPs i not analysed deeply for practical use.

The fitness function F based on this modification is

$$F(\vec{X_i}, O_i) = -\sqrt{\frac{\sum_{i \in I}\left(\frac{1}{1+e^{-Raw(\vec{X_i})}} - O_i\right)^2}{|I|}} + d + 1$$

where $\vec{X_i}$ is an input case included in a set of I and O_i is the outcome matching to an input i. $d + 1$ is a constant in terms of the depth of the tree. The internal function $Raw(\vec{X_i})$ returns the evaluated value of a tree individual with the ith input case. Maximising this fitness function up to 4, we can obtain individuals minimising RMSE.

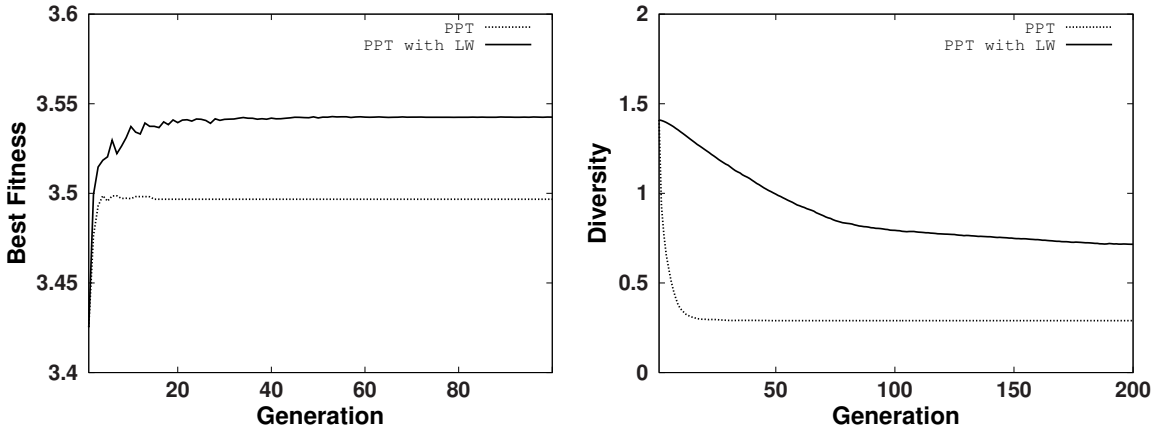
EDA-GP Systems. For overall experiments, we used a basic PPT-based EDA-GP system similar to PIPE. Its detail parameters are shown in the table 2.

Table 2. Parameter Settings for the EDA-GP System

Genotype	tree	Selection	Truncation
Max.Depth	3	Ratio	0.3
Operators	$+, -, \times, /$	Sampling	Ancestral
Terminals	$X1, \ldots, X7, ERC$	Update	Max. Likelihood
Population	30	Learning Ratio	0.9
Generations	200	Runs	30

5. Results

Performance. In figure 2, mean best fitness of two PPT systems is shown, averaged across 30 runs at each generation. In this graph, the ordinary PPT-based system starts from 3.425 and converges to 3.497±0.18. After applying LW, it converges 3.543±0.002. Corresponding RMSE of the former fitness is 0.503 and the latter is 0.457, since the fitness function is sum of RMSE and constant depth factor. In one-tailed *Welch's t-test* for the results, p-value is 0.086. To evaluate problem complexity, we also evaluated a simple model assuming that all cases are true or false. Achievable RMSE is 0.548 in the given data composed of 1,422 false among 4,738 samples.



(a) Performance (b) Diversity

Fig. 2. Performance and Diversity of the PPT-based EDA-GPs

Diversity. To evaluate behaviour changes of EDA-GP systems, we measured diversity of PPT models through a metric defined by Shapiro[? ?] for EDAs. This metric is a trace of covariance matrix composed of probability of all independent random variables used in PPT, whose detailed explanation is shown in papers[?] . In figure 2, diversity of the PPT-based system decreases quickly in early 20 generations, and then converges to 0.290±0.224. After applying LW, its converging point increased to 0.716±0.103 and dropped slower than the ordinary system.

6. Discussions

The PPT-based EDA-GP systems improve the prediction of arthritis problem, and applying LW method is confirmed to ameliorate their performance and diversity. In the result, it shows more accurate prediction by 4.6% of maximum RMSE compared to the ordinary PPT-based system and maintains diverse probability distribution for search. In the ordinary PPT-based system, it still improves accuracy than the simplest hypothesis.

Beyond confirming the effectiveness of applying LW method, it may be possible to find better local optima by applying a variety of advanced techniques in selection, sampling and learning. We leave such accuracy improvement as future research and possible applications to more critical medical problems as well.

7. Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(Project No.2012–004841). The ICT at Seoul National University provided research facilities for this study.